
Case-Only Design to Measure Gene-Gene Interaction

by Quanhe Yang, Muin J. Khoury, Fengzhu Sun, W. Dana Flanders

Abstract

The case-only design is an efficient and valid approach to screening for gene-environment interaction under the assumption of the independence between exposure and genotype in the population. In this paper, we show that the case-only design is also a valid and efficient approach to measuring gene-gene interaction under the assumption that the frequencies of genes are independent in the population. Just as the case-only design requires fewer cases than the case-control design to measure gene-environment interaction, it also requires fewer cases to measure gene-gene interactions.

Genetic studies have evolved from simple analysis of single genes to include more sophisticated analysis of complex traits, an evolution that

parallels an increasing recognition of the role of gene-environment interactions in disease etiology (1). Genetic factors probably contribute to virtually every human disease, conferring susceptibility or resistance, or influencing interaction with environmental factors. In addition to their use in studies of examining gene-environment interactions in human complex traits, genetic studies have increasingly been used to examine the effects of gene-gene interactions in disease etiology. Examples include studies of the interaction of 5,10 methylenetetrahydrofolate reductase (MTHFR), and cystathionine-beta synthase (CBS) in determining risk for neural tube defects (NTDs) (2-5); a study of the interaction of apo E/apo C-I, angiotensin-converting enzyme (ACE), and MTHFR, in determining risk for coronary artery disease (CAD) (6); and a study of factor V Leiden and MTHFR, in determining genetic susceptibility to preeclampsia (7). Many other studies of the contributions of genetic factors in human diseases suggest that gene-gene interactions may play an important role in the etiology of the diseases studied (8-12). As the Human Genome Project

provides further information on all human genes, the studies of gene-gene interactions will play an increasingly important role in the search for the causes of human diseases.

Previous work has shown that the case-only design is an efficient and valid approach to screening for gene-environment interaction under the assumption of independence between exposure and genotype in the population (13, 14). In this paper, we demonstrate that the case-only design is also a valid approach to measuring the effects of gene-gene interactions, assuming that the genes under study are not in linkage disequilibrium. Like studies of gene-environment interactions, studies of gene-gene interactions require fewer case subjects to measure gene-gene interactions if they use a case-only design rather than a case-control design.

METHODS

While there is no unified definition of gene-gene interaction, we may broadly define the gene-gene interaction as the effects of one or more genes in

determining the occurrence of the diseases are modified by the presence or absence of another gene or genes. In the study of gene-gene interaction, the primary interest is to assess the proportion of disease among those who are jointly exposed is due to the interaction of the exposure to the two or more gene variants.

For the purpose of demonstration, we assumed two disease susceptibility genes (gene 1 and gene 2) non-linked at two loci with gene frequencies of P_1 and P_2 in the population. Each of these disease susceptibility genes has two allelic variants (susceptible and non-susceptible) that follow an autosomal dominant inheritance pattern. If the two genes under study are on the same chromosome, we assume that they are not in linkage disequilibrium in the population at risk. We also assume the existence of background risk unrelated to either gene.

For the two diallelic genes, let the first subscript i indicate that the variant of gene 1 is present (1) or absent (0), and second subscript j indicate the present (1) and absent (0) of the variant of gene 2. Let p_{ij} denote the proportion of the

population who have the variant of gene 1 at level i (i.e., i=1 for present, i=0 for absent) and the variant of gene 2 at level j. Let R_{ij} indicate the risk associated with the combinations of present and absent of the variants of gene 1 and gene 2, e.g., R_{10} indicates the risk of disease for persons having gene 1 variant alone and R_{01} indicates the risk of disease for persons having gene 2 variant alone.

Table 1 shows the distribution of the number of cases expected to arise during follow-up of a “fixed” population in terms of genes frequencies in the population and risks associated with the combination of present and absent of the gene variants. We can construct a 2-by-2 table using cases only by the presence and absence of the gene 1 and gene 2 variants (Table 2). From this table of representative subset of cases, the case-only cross-product (χ^2_{co}) is:

$$\chi^2_{co} = \frac{ad}{bc} = \frac{(P_{11} \cdot N \cdot R_{11})(P_{00} \cdot N \cdot R_{00})}{(P_{10} \cdot N \cdot R_{10})(P_{01} \cdot N \cdot R_{01})} = \frac{(P_{11} \cdot R_{11})(P_{00} \cdot R_{00})}{(P_{10} \cdot R_{10})(P_{01} \cdot R_{01})} \quad (1)$$

Let a “.” in the subscript refer to the marginal (overall) frequency of the genes in the population, so that $p_{1.}$ refers to the marginal frequency of gene 1 variant, $p_{.1}$

refers to the marginal frequency of gene 2 variant, $p_{.0}$ refers to the proportion of the population without mutant allele of gene 1, and $p_{0.}$ refers to the proportion of the population without mutant allele of gene 2. We assume that two genes are distributed independently in the population and are not in linkage disequilibrium, then $p_{11} = p_{1.} \cdot p_{.1}$, $p_{10} = p_{1.} \cdot p_{.0}$, $p_{01} = p_{0.} \cdot p_{.1}$, and $p_{00} = p_{0.} \cdot p_{.0}$.

Substituting these marginal frequencies into equation 1 for case-only φ_{co} , we have:

$$\varphi_{co} = \frac{(P_{1.} \cdot P_{.1} \cdot R_{11}) (P_{0.} \cdot P_{.0} \cdot R_{00})}{(P_{1.} \cdot P_{.0} \cdot R_{10}) (P_{0.} \cdot P_{.1} \cdot R_{01})} = \frac{R_{11} R_{00}}{R_{10} R_{01}} \quad (2)$$

If we define the risk ratios as: $RR_{11} = R_{11}/R_{00}$, $RR_{10} = R_{10}/R_{00}$, and $RR_{01} = R_{01}/R_{00}$,

φ_{co} can be expressed in terms of risk ratios as: $\varphi_{co} = RR_{11} / RR_{10} \cdot RR_{01}$.

Assuming that the genes under study are not in linkage disequilibrium, if the effects for the two genes conform to a multiplicative relation, then the case-only φ_{co} in a representative sample of cases only should equal unity, i.e., $RR_{11} / RR_{10} \cdot RR_{01} = 1$. When the case-only φ_{co} departs from unity, either the population

frequencies of the genes are not independent or the gene-specific event rates do not conform to a multiplicative relation of joint effect. Under the assumption of independent gene frequencies in the population, the case-only θ_{co} provides an estimate of the ratio of the joint effect (RR_{11}) divided by the product of the individual effects of each gene alone (i.e., $\theta_{co} = RR_{11} / (RR_{10} \cdot RR_{01})$), which can be regarded as effect measure modification of risk ratio on a multiplicative scale or a gene-gene interaction of the risk ratio.

Thus, for assessing gene-gene interaction in the etiology of a disease, investigators can use a case-only design if the two genes under study are in linkage equilibrium. Although we present the case-only θ_{co} in the context of gene-gene interaction, the algebraic relation involved applies to any two factors that are distributed independently in a population whose cases are identified or sampled in proportion to their occurrence.

The approach proposed here differs from that proposed by other authors who used a logistic model to measure gene-environment interaction with the

case-only design (13, 14). We have shown that the cross-product term (ϕ_{co}) in a case-only 2-by-2 table measures the departure from the multiplicative joint effects of risk ratios, but not odds ratios. Our results show that if the cross-product in a case-only design is unity, then the risk ratios multiply. That is, if $\phi_{co} = 1$, then,

$$R_{11} = \frac{R_{10}R_{01}}{R_{00}} \quad (3)$$

Thus, use of the cross-product in a case-only design to measure gene-gene interaction reflects departure from multiplicativity of risk ratios.

On the other hand, Piegorsch et al.(13) formulated disease risk using a logistic model to measure gene-environment interaction. They assumed rare disease and a logistic model, and showed that the cross-product in a case-only design reflected departure from multiplicativity of odds ratios. In other words, if there is no interaction,

$$\frac{R_{11}}{(1 \& R_{11})} \cdot \frac{(\frac{R_{10}}{1 \& R_{10}})(\frac{R_{01}}{1 \& R_{01}})}{(\frac{R_{00}}{1 \& R_{00}})} \quad (4)$$

where the first subscript (i = 0 or 1) indicates the presence (1) or absence (0) of an arbitrary disease susceptibility gene, and second subscript indicates the presence (1) and absence (0) of environmental exposure for the study of gene-environment interaction. This expression has the same form as equation (3), except that odds ratios appear in place of risk ratios.

For a rare disease, risk ratios approximate odds ratios so that our results imply that the cross-product measures departure from a multiplicative relation of odds ratios, as shown by Piegorsch et al.(13). Our results, however, also show that the cross-product (χ^2_{co}) remains a valid measure of departure from multiplicativity of risk ratios even if the disease is not rare.

EXAMPLE

Ramsbottom et al. (5) studied the relation between specific MTHFR and

CBS polymorphism and NTD risk with 127 case and 430 control subjects in an Irish population. Botto et al. (15) re-analyzed data derived from that study using a 2-by-4 table with estimated gene (MTHFR and CBS) frequencies in the population. Both found that, compared with subjects with neither mutation, the NTD risk was 2.1 times higher among those who had the MTHFR mutation only (95% CI, 1.1-3.9), 0.8 times higher among those who had CBS mutation only (95% CI, 0.4-1.4), and 5.2 times higher among those who had both mutations (95% CI, 1.4-21.2). These results indicate an odds ratio of 3.1 (95% CI=0.8-13.1) for gene-gene interaction ($OR_{int} = 5.2 / (0.8 * 2.1)$), the factor by which the odds ratio for those exposed to gene 1 and gene 2 is different from the multiplied effect of each gene alone. Using the case-only design, we estimate an interaction risk ratio of 2.0 (95% CI 0.6-6.0) for gene-gene interaction, also implying interaction.

Like studies of gene-environment interaction using case-only design (16), studies of gene-gene interaction also require fewer case subjects if a case-only design is used than if a case-control design is used. Furthermore, because the

risk for NTD is low, our results are similar to those that would be derived from a logistic model.

DISCUSSION

Our results show that the case-only design is a valid and relatively efficient approach to measuring gene-gene interaction under the assumption that gene frequencies in the population are independent (e.g., linkage equilibrium or independent assortment). With the rapid progress of the Human Genome Project, which will provide further information on human genes, studies of gene-environment and gene-gene interactions will play an increasingly important role in determining the etiology of complex human diseases. As we have shown, the case-only design can be an effective means of conducting these studies.

It is important to point out that the definition and measurement of interaction has been a subject for debate in the epidemiologic literature (17-27). A major distinction is between the concepts of statistical and biological interaction.

Statistical interaction occurs if the effects of two or more risk factors is not additive on an arbitrary scale of measurement. This concept of interaction has been criticized because it ignores the concept of biological interaction, and is inherently arbitrary and model-dependent (22, 26-27). Biological interaction refers to the coparticipation of two risk factors in the same causal mechanism for the disease development (18, 27). Complete absence of biologic interaction can, under certain conditions, imply additivity of risk differences (27). In measuring biological interaction, the primary interest is not to conduct the statistical modeling, but to assess the proportion of disease among those who are jointly exposed to both factors that may be due to the interaction of the two exposures. Many studies have suggested that for addressing public health concerns regarding disease frequency reduction, biologic interaction, i.e., assessing deviations from additivity, are most relevant (18, 20, 24, 27).

As noted before, the cross-product (χ^2_{co}) in a case-only design measures the ratio of RR_{11} to $RR_{01} \cdot RR_{10}$, a measure of departure from risk ratio multiplicativity.

On the other hand, the state of no interaction on an additive scale implies: $RR_{11} =$

$RR_{01} + RR_{10} - 1$. Unfortunately, the risk ratios associated with each gene alone

(RR_{01} or RR_{10}) cannot be estimated without employing proper control population.

It is not difficult to show that for two genes, $(RR_{01} \cdot RR_{10}) = (RR_{01} + RR_{10} - 1)$ if the

effect of one of the gene (RR_{01} or RR_{10}) = 1 or both RR_{01} and $RR_{10} = 1$. Under these

circumstances, a departure from multiplicativity is equivalent to a departure from

additivity. In this situation, the case-only design can indicate departure from

additivity, which, under certain conditions, can be derived from the biologic null

of independent action (27). Furthermore, if both genes confer increased risk,

$RR_{01} \cdot RR_{10}$ is greater than $(RR_{01} + RR_{10} - 1)$, which implies that the measured

departure from multiplicativity must reflect an even greater departure from

additivity. The differences between $RR_{01} \cdot RR_{10}$ and $(RR_{01} + RR_{10} - 1)$ approach zero

as the effects of both genes acting alone approach one. Thus, the case-only design

can be used, under certain conditions, to indicate departure from additivity, and

in turn, biologic interaction.

There are several limitations, however, to using the case-only design to measure gene-gene interactions. The most important limitation is that the case-only design does not measure departure from additivity, it measures departure from the multiplicative joint effects of two risk ratios. Under certain conditions, i.e., RR_{01} or $RR_{10} = 1$ or both RR_{01} and $RR_{10} = 1$, a departure from multiplicativity equals a departure from additivity. The investigators should keep in mind this important limitation when using case-only design to measure the effects of gene-gene interaction. Second, the genes under study must be in linkage equilibrium, assort independently, or otherwise have independent genes frequencies in the population being studied. Linkage disequilibrium between any genes being studied can invalidate a case-only design to measure gene-gene interaction. In studies of most disease susceptibility genes in which the case-only design would be applied, we suspect that the investigators would know where the genes are located. Genes on different chromosomes are unlikely to be correlated. The probability of correlations (linkage disequilibrium) between genes on the same

chromosome will increase, especially for those that are physically close to each other. Investigators will have to interpret results of case-only studies cautiously or use another approach when examining the gene-gene interactions for genes that may be in linkage disequilibrium. Third, the case-only design cannot estimate the risk associated with each gene alone (R_{10} or R_{01}). The effect of each gene alone could be estimated from a case-control study, a case-parental control study or other types of studies (28-29). Finally, the effects of population stratification may also invalidate the results of a case-only study. For example, if two gene frequencies occur together commonly in a particular ethnic population, and this population also has high risk for the disease, the effect of gene-gene interaction can be overestimated. One can try to restrict the analysis to avoid that particular ethnic population.

Investigators are increasingly searching for gene-gene interactions in human complex traits. With the rapid progress in molecular technology and the Human Genome Project, there will be increased interest in searching for the

effects of gene-environment and gene-gene interactions in disease etiology. The case-only design is a useful tool with which to rapidly screen for these interactions.

ACKNOWLEDGMENTS

We thank Dr. Lorenzo D. Botto for providing with us the example data and Dr. J. David Erickson for his helpful comments, as well as one of the anonymous reviewers for his/her constructive comments and suggestions.

REFERENCES

1. Khoury MJ. Genetic epidemiology. In: Rothman KJ, Greenland S. eds. *Modern Epidemiology*. Philadelphia, PA: Lippincott-Raven Publishers, 1998:609-622.
2. Whitehead AS, Gallagher P, Mills JL, Kirke PN, Burke H, Molloy AM, Weir DG, Shields DC, Scott JM. A genetic defect in 5,10 methylenetetrahydrofolate reductase in neural tube defects. *QJM* 1995;88(11):763-766.
3. Van der Put NM, Steegers-Theunissen RP, Frosst P, Trijbels FJ, Eskes TK, van den Heuvel LP, Mariman EC, den Heyer M, Rozen R, Blom HJ. Mutated methylenetetrahydrofolate reductase as a risk factor for spinal bifida. *Lancet* 1995;346(8982):1070-1071.
4. Ou CY, Stevenson RE, Brown VK, Schwartz CE, Allen WP, Khoury MJ, Rozen R, Oakley GP Jr, Adams MJ Jr. 5,10 methylenetetrahydrofolate reductase genetic polymorphism as a risk factor for neural tube defects. *Am J Med Genet* 1996;63:610-614.
5. Ramsbottom D, Scott JM, Molloy A, Weir DG, Kirke PN, Mills JL, Gallagher PM, Whitehead AS. Are common mutations of cystathionine beta-synthase involved in the aetiology of neural tube defects? *Clin Genet* 1997;51:39-42.
6. Galinsky D, Tysoe C, Brayne CE, Easton DF, Huppert FA, Denning TR, Paykel ES, Rubinsztein DC. Analysis of the apo E/apo C-I, angiotensin converting enzyme and methylenetetrahydrofolate reductase genes as candidates affecting human longevity. *Atherosclerosis* 1997;129:177-183.
7. Grandone E, Margaglione M, Colaizzo D, Cappucci G, Paladini D, Martinelli P, Monranaro S, Pavone G, Di Minno G. Factor V Leiden, C>T MTHFR polymorphism and genetic susceptibility to preeclampsia. *Thromb & Haemost.* 1997;77:1052-1054.
8. Jarvik G, Larson EB, Goddard K, Schellenberg GD, Wijsman EM. Influence of apolipoprotein E genotype on the transmission of Alzheimer disease in a community-based sample. *Am J Hum Genet* 1996;58:191-200.
9. MacLeod S, Sinha R, Kadlubar FF, Lang NP. Polymorphism of CYP1A1 and GSTM1 influence the in vivo function of CYP1A2. *Mutat Res* 1997;376:135-142.
10. Beales PL, Kopelman PG. Obesity genes. *Clin Endocrinol* 1996;45:373-378.
11. Williams RR, Hunt SC, Hopkins PN, Wu LL, Hasstedt SJ, Berry TD, Barlow GK, Stults BM, Schumacher MC, Ludwig EH. Genetic basis of familial dyslipidemia and hypertension: 15-year results from Utah. *Am J Hypertens* 1993;6(11 pt 2):319S-27S.
12. Davignon J, Roy M. Familial hypercholesterolemia in French-Canadians: taking advantage of the presence of a 'founder effect'. *Am J Cardiol* 1993;72:6D-10D.
13. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med* 1994;13:153-162.
14. Begg CB, Zhang ZF. Statistical analysis of molecular epidemiology studies employing case-series. *Cancer Epidemiol Biomarkers Prev* 1994;3:173-175.
15. Botto LD, Mastroiacovo P. Exploring gene-gene interactions in the etiology of

neural tube defects. *Clin Genet* 1998;53:456-459.

16. Yang QH, Khoury MJ, Flanders WD. Sample size requirements in case-only designs to detect gene-environment interaction. *Am J Epidemiol* 1997;146:713-720.

17. Koopman JS. Causal models and sources of interaction. *Am J Epidemiol* 1977;106:439-44.

18. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. *Am J Epidemiol* 1980;112:467-70.

19. Siemiatycki J, Thomas DC. Biological models and statistical interactions. *Int J Epidemiol* 1981;10:383-87.

20. Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research*. New York: Van Nostrand Reinhold, 1982.

21. Miettinen OS. Causal and preventive interdependence: elementary principles. *Scand J Work Environ Health* 1982;18:159-68.

22. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. *Stat Med* 1983;2:243-51.

23. Greenland S, Poole C. Invariants and noninvariants in the concept of interdependent effects. *Scand J Work Environ Health* 1988;14:125-29.

24. Pearce N. Analytical implications of epidemiological concepts of interaction. *Int J Epidemiol* 1989;18:976-80.

25. Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. *J Clin Epidemiol* 1991;44:221-32.

26. Greenland S. Basic problems in interaction assessment. *Environ Health Perspect* 1993;101(suppl 4):59-66.

27. Rothman KJ, Greenland S. *Modern Epidemiology*. Philadelphia, PA: Lippincott-Raven Publishers, 1998.

28. Khoury MJ, Flanders WD. Non-traditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! *Am J Epidemiol* 1996;144:207-213.

29. Yang QH, Khoury MJ. Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. *Epidemiol Rev* 1997;19:33-43.

TABLE 1. The expected distribution of cases for gene-gene interaction analysis by gene frequencies in the population and risks associated with these gene variants in a case-only design

Gene 1 variant	Gene 2 variant	Gene frequencies in population	Risk associated with genes	No. of expected cases
+	+	p_{11}	R_{11}	$p_{11} \cdot R_{11} \cdot N$
+	-	p_{10}	R_{10}	$p_{10} \cdot R_{10} \cdot N$
-	+	p_{01}	R_{01}	$p_{01} \cdot R_{01} \cdot N$
-	-	p_{00}	R_{00}	$p_{00} \cdot R_{00} \cdot N$

Where

p_{11} = proportion of population who have both gene variants.

p_{10} = proportion of population who have gene 1 variant only.

p_{01} = proportion of population who have gene 2 variant only.

p_{00} = proportion of population who have neither mutant genes.

R_{11} = disease risk of having both gene variants (gene 1 variant = 1 and gene 2 variant = 1).

R_{10} = disease risk of having gene 1 variant alone (gene 1 variant = 1 and gene 2 variant = 0).

R_{01} = disease risk of having gene 2 variant alone (gene 1 variant = 0 and gene 2 variant = 1).

R_{00} = back ground disease risk (gene 1 variant = 0 and gene variant 2 = 0).

TABLE 2. Case-only 2-by-2 table classified by the presence and absence of gene 1 variant and gene 2 variant

Gene 1 variant	Gene 2 variant	
	+	-
+	a	b
-	c	d

Where $a = p_{11} \cdot R_{11} \cdot N$
 $b = p_{10} \cdot R_{10} \cdot N$
 $c = p_{01} \cdot R_{01} \cdot N$
 $d = p_{00} \cdot R_{00} \cdot N$

$\theta_{co} = ad/bc$

TABLE 3. Estimated odds ratios of MTHFR and CBS gene variants for NTD risk

MTHFR	CBS	Cases	Controls	OR	95% CI
+	+	7	5	5.2	1.4-21.2
+	-	19	34	2.1	1.1-3.9
+	-	16	76	0.8	0.4-1.4
+	-	85	315	1.0	-

Derived from Botto et al.(15).

In a case-control design, the odds ratio for MTHFR (+) and CBS (+) = 5.2, and gene-gene interaction = 3.1 (95% CI, 0.8-13.1) since R_{10} (MTHFR) = 2.1, R_{01} (CBS) = 0.8. The estimation of effect of gene-gene interaction using a case-only design, $\theta_{co} = (7 \times 85) / (19 \times 16) = 2.0$ (95% CI, 0.6-6.0).